The bait compatibility index: a computational approach towards bait selection for interaction proteomics experiments

Sudipto Saha[1], Parminder Kaur[1] and Rob Ewing[1,2]

[1]Center for Proteomics and Bioinformatics, [2]Department of Genetics, School of Medicine, CWRU, Cleveland, OH.

Abstract

Background: Yeast two-hybrid (Y2H) and affinity-purification mass-spectrometry (AP-MS) are two commonly used techniques for large-scale detection of protein interactions. Y2H identifies binary physical interactions by detecting reconstitution of a split transcription factor via activation of a reporter gene. On the other hand, AP-MS captures protein complexes near physiological conditions, and combines the specificity of antibody-based protein purification with the sensitivity of mass spectrometry. These techniques provide fundamentally different views of the protein interactome, and it is not clear what the specific biases in each technique really are**. Methods:** Here, we systematically study these biases and generate a novel score, the bait compatibility index, that can be used to select baits according to their compatibility with each technique. First, naïve Bayesian models were created based on sequence and annotated features and abundance of bait proteins using yeast interaction proteomics data to predict experimental outcomes. Second, we identified significantly enriched terms across successful and unsuccessful baits in Y2H and AP-MS. **Results:** We observe an accuracy of 71.25% using the optimum four features (post-translational modifications, sub-cellular location, pathway and abundance) in the AP-MS dataset, whereas Y2H experimental outcomes could be predicted to an accuracy of 63.38% using only two features (GO molecular function and pathway). A set of 391 significant annotation terms (p-value <0.05) were identified as over- represented in the AP-MS or Y2H methods. **Conclusion:** We demonstrate that significant bias in interaction proteomics datasets can be attributed to bait features by in-depth analysis of serine/threonine phosphatase and peroxisome baits. We also show that the yeast models may also be applied to human datasets, and provide a valuable predictor of the suitability of baits for interaction proteomics experiments.

Introduction

Protein interaction mapping is an important area of proteomics and large scale studies on yeast, nematode and human were performed using two major technologies: Y2H and AP-MS [1,2,3,4]. AP-MS identifies protein complexes, whereas Y2H detects binary interactions. Networks derived from Y2H and AP-MS data have different topologies, and the fundamental reasons for bait to be successful/unsuccessful are not fully clear.

A dataset of 4135 genes used in both methods in yeast was used in our study for developing naïve Bayesian models. Seven features were selected for annotating these genes [Table 1].The frequency of terms for successful/unsuccessful baits were computed. The work flow for the data analysis in our study is shown in Fig 1. The naïve Bayesian model calculates posterior probabilities for a given hypothesis (successful/unsuccessful bait) assuming that the features that describe instances are conditionaly independent [5]. The combination of features were used to find an optimum model based on performance measure such as sensitivity, specificity, accuracy and ROC. Each bait in our study was assigned a bait compatibility index for Y2H and AP-MS methods, which represents the log likelihood of a successful/unsuccessful outcome.

To identify whether the annotated terms for each features were biased in one method or the other, a two-tailed Fisher's exact test was performed by constructing a 2x2 contingency table for the frequency for each term across successful and unsuccessful classes in Y2H and AP-MS datasets.

References

1. Yu H et al., High-quality binary protein interaction map of the yeast nteractome network. Science. 2008;322(5898):104-10.
2. Krogan NJ et al., Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006; 440(7084):637-43.
3. Rual JF et al., Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005;437(7062):1173-8.
4. Ewing RM et al., Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol. 2007;3:89.
5. Mitchell, Tom. "Bayes Learning" Machine Learning. New York:McGraw-Hill. 1997; pp-154-200
6. Ghaemmaghami S et al., Global analysis of protein expression in yeast. Nature. 2003;425(6959):737-41.

Rationale

Bait Compatibility Index (BCI): Bait selection is an important step in AP-MS and Y2H based methods to identify protein interaction partners. Large-scale studies in the yeast interactome show that approximately 50% the bait proteins were successful in identifying interacting partners in both methods. This encouraged us to study those bait proteins to investigate whether any features or terms annotated to the bait proteins have any underlying reasons to become successful or unsuccessful. To measure whether the confidence for a bait protein is likely to be successful/unsuccessful, we came up with BCI scores. BCI is defined as a log likelihood ratio of posterior probability of successful and unsuccessful outcomes for a given protein. This allows the researchers to prioritize the baits *in silico* and this scoring method is better than the random selection of baits.

Features selection: Three criteria were applied in the selection of the final seven features used in the model. 1) Features were selected that are known as important mediators of protein interactions (e.g. domain/motifs, molecular function). 2) Features which are well annotated across genome-wide datasets. 3) The sets of features that shows minimal dependence upon each other were evaluated so that they can be appropriately modeled using the naïve Bayesian model.

Naïve Bayesian (NB): Bayesian models allow us to combine dissimilar types of data (i.e. numeric and categorical) and converting them to a common probabilistic framework. A naïve Bayesian predictor was used to prioritize ranking of bait proteins correlated to be successful by integrating sequence and annotated features of yeast proteins. Evaluation of the importance of features and terms in the neural network or the support vector machine is not possible from biological insight. Thus we have focused on NB for its best accuracy and simplicity.

Feature	Description	Source
F1	Post-translational modification	UniProtKB
F2	Sub-cellular location	UniProtKB
F3	Prosite motifs	GenomeNet-Kegg
F4	Gene Ontology Biological process	UniProtKB
F5	Gene Ontology Molecular function	UniProtKB
F6	Pathway	GenomeNet-Kegg
F7	Abundance	[Ref. 6]

Table 1. Annotation features (F1-F7) and sources (UniProtKB was release 14.5, Kegg release 48.0) used to construct feature vectors for each bait.

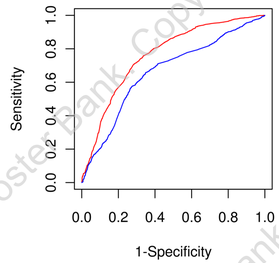

Figure. The ROC plot of AP-MS (red) and Y2H(blue) optimum models of yeast interaction datasets. Four features (F1: PTMs; F2: Sub-cellular location; F6: Pathway; F7: Abundance) combination achieved maximum accuracy in the AP-MS dataset where as two features (F5: GO Molecular function; F6: Pathway) achieved maximum accuracy in the Y2H dataset. The areas under the curve for AP-MS and Y2H models are 0.76 and 0.66 respectively.

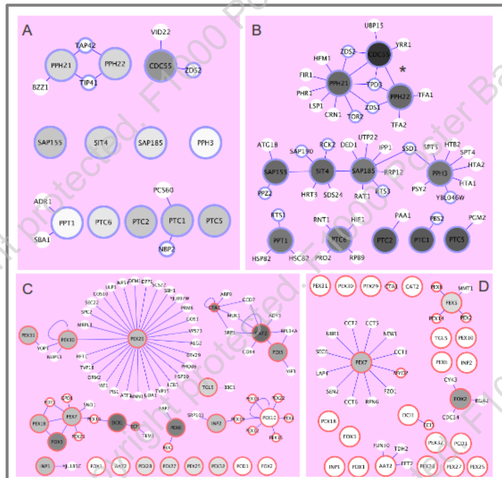

Figure. Comparative analysis of selected Y2H and AP-MS yeast networks. (A,B) A set of 12 serine/threonine phosphatase baits, tested using Y2H (A) and AP-MS (B) are shown with their associated interaction networks. * Network corresponding to serine/threonine phosphatase 2A (PP2A). The underlying serine/threonine phosphatase network is revealed with AP-MS but not Y2H. (C,D) Peroxisome-annotated baits show a rich interaction network from Y2H networks (C) in comparison to AP-MS (D). (E) Legend: Baits are represented as large nodes, prey as small nodes. In each case, baits annotated as serine threonine phosphatases or functioning in the peroxisome are outlined in color. Preys outlined in color in (A,B) represent phosphatase components or proteins with phosphatase regulatory activity. Preys outlined in color in (C,D) represent proteins localized to the peroxisome. The strength of the bait score, assessing the compatibility of each bait with each technology is represented by shaded (white-gray-black) nodes.

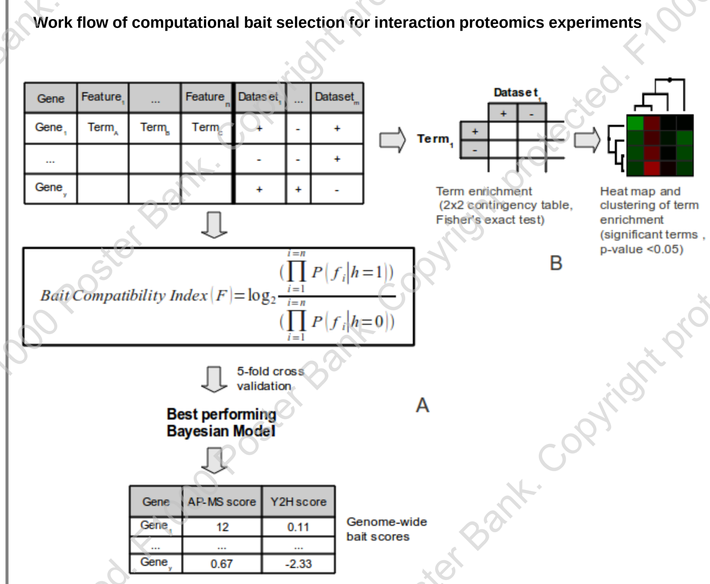

Figure 1. Overview of data processing work-flow. (**A**) Training and testing of Bayesian model to predict success according to features of each bait. (**B**) Identification of annotation terms significantly enriched in successful and unsuccessful baits

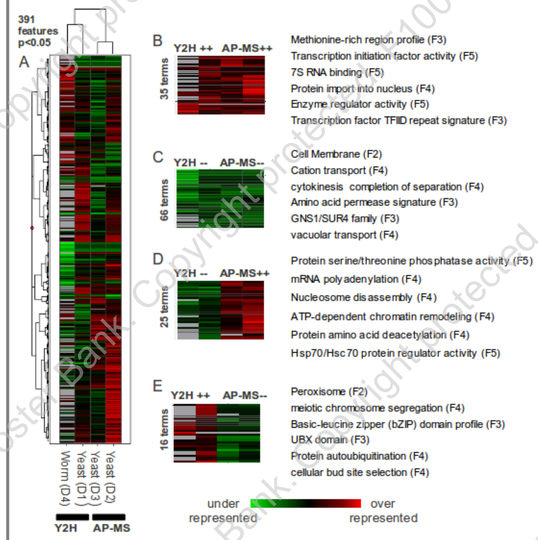

Figure. Hierarchical clustering enables visualization of patterns of enrichment of bait features across studies and technologies. (A) Heat-map showing all 391 significant terms (Fisher's exact test, p-value <0.05). Subsets of terms showing: (B) overall enrichment in successful baits, (C) overall enrichment in unsuccessful baits, (D) enriched in successful AP-MS baits and enriched in unsuccessful Y2H baits, (E) enriched in unsuccessful AP-MS baits and enriched in successful Y2H baits. Example terms are shown for each class at far right.

Summary

- We analysed the differences between the two major technologies, Y2H and AP-MS for mapping protein interaction data and came up with BCI scores for each bait protein based on a naïve Bayesian model.
- We studied features and terms that relate to the successful/unsuccessful identification in both methods.
- The predictive power of the AP-MS model is higher when compared to the Y2H model.
- Unlike Y2H, the AP-MS based model is dependent on features like sub-cellular location and abundance.

Acknowledgments

We thank Gaurav S. J. B Rana for assistance with statistical testing. RE acknowledges start-up funds from the Cleveland Foundation and Center for Proteomics and Bioinformatics at Case Western Reserve University.